Introduction & Data Examples

Alexander McLain

August 18, 2025

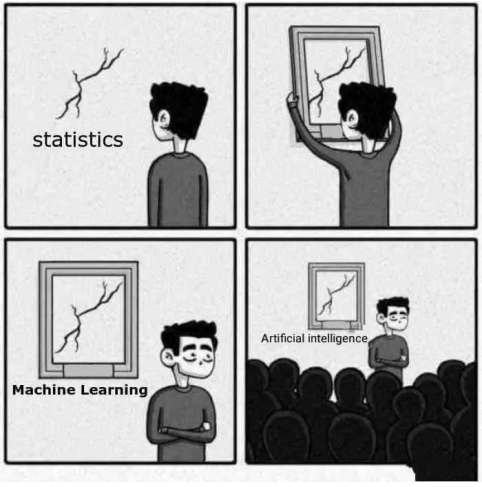


Figure: Great figure that I don't 100% agree with. Original comic by sandserif

Daniela Witten
@daniela_witten

"When we raise money it's AI, when we hire it's machine learning, and when we do the work it's logistic regression."

(I'm not sure who came up with this but it's a gem 💎)

2:50 PM · Sep 26, 2019

AI vs ML

Artificial intelligence (AI) and machine learning (ML) are closely related fields, but they are not the same.

- **AI:** broad field that involves creating systems or machines that can perform tasks that typically require human intelligence. These tasks include reasoning, problem-solving, understanding natural language, perception, and decision-making.
  - **Examples:** virtual assistants like Siri and Alexa, ChatGPT, chess-playing computers, and autonomous vehicles.

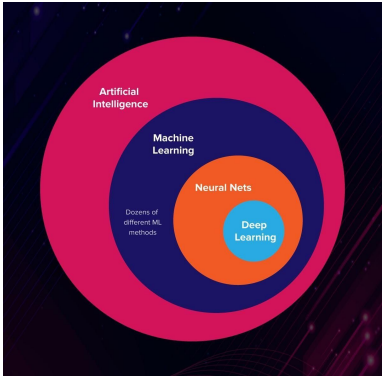
AI vs ML

Artificial intelligence (AI) and machine learning (ML) are closely related fields, but they are not the same.

- **AI:** broad field that involves creating systems or machines that can perform tasks that typically require human intelligence. These tasks include reasoning, problem-solving, understanding natural language, perception, and decision-making.
  - **Examples:** virtual assistants like Siri and Alexa, ChatGPT, chess-playing computers, and autonomous vehicles.
- **ML** is a subset of AI that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data.
  - **Examples:** spam filters in email systems, recommendation engines on e-commerce sites, image and speech recognition systems, and predictive analytics.

AI vs ML

- AI is a broader concept that includes ML as one of its components. ML is a critical component of AI, but AI is not limited to ML.
- Many modern AI applications, such as image recognition, natural language processing (i.e., ChatGPT), and game playing, rely heavily on ML techniques to function effectively.
- ML is focused on the development and application of algorithms that can learn from data.
- Neural Networks are a common ML method.



Machine learning

- Machine learning has increasingly been applied in the field of public health over the past few years.
- This has offered novel solutions to a range of complex problems.
- The next few slides give some general examples of how machine learning is being used in public health.

Examples of ML

1. Disease Surveillance:
   - Outbreak Prediction: Machine learning models can analyze diverse datasets, including climate data, social media posts, and traditional health reports, to predict disease outbreaks. For instance, predicting flu or dengue outbreaks based on weather patterns and previous case reports.
   - Outbreak Prediction with media.
   - Latent infectious disease monitoring.
   - Disease Monitoring: Algorithms can process vast amounts of data quickly to detect anomalies or outbreaks in real time, allowing for quicker public health responses.
   - Disease Monitoring.
   - Resource Allocation.

Examples (cont.)

2. Disease Diagnosis and Risk Prediction:
   - Models can analyze electronic health records (EHRs) to predict high-risk individuals for certain diseases, ensuring timely interventions.
   - Clustering EHR data.
   - Diagnostic tools, particularly using medical imaging data, can be developed to detect diseases like tuberculosis, pneumonia, or diabetic retinopathy.
   - Review
3. Genomic Data Analysis:
   - Machine learning can handle vast genomic datasets, leading to insights related to disease susceptibility, personalized treatments, or understanding disease mechanisms at the genetic level.
   - Predicting Obesity Based on Genomic Data
   - Genome review.

Examples:

4. Treatment Personalization:
   - Machine learning can help tailor treatments to individuals, understanding which treatments might work best for which individuals based on genetics, lifestyle, and other factors.
   - Treatment Personalization.
   - Treatment Personalization with reinforcement learning
   - Review.
5. Drug Discovery and Repurposing:
   - Using machine learning, patterns in large datasets can be recognized, leading to insights for drug discovery or finding new uses for existing drugs.
   - Don't be bitter.
   - Review.

Examples (cont.)

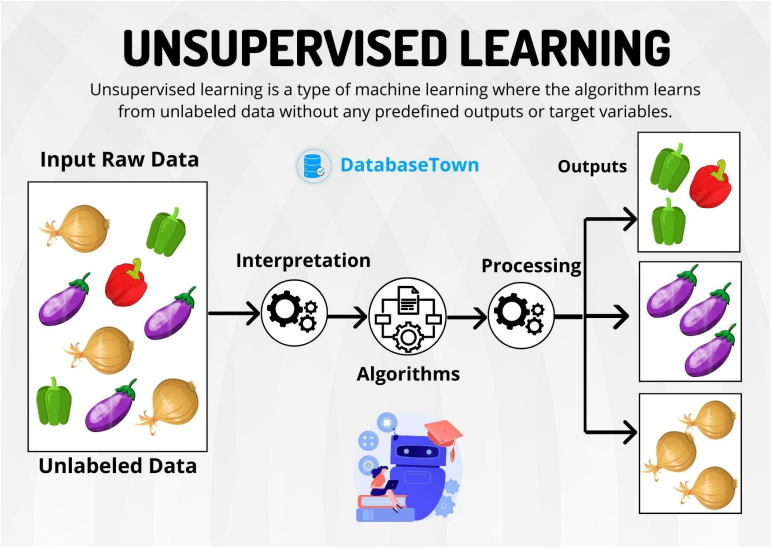
6. Health Behavior and Information Spread:
   - Analyzing social media data can offer insights into health behaviors, misconceptions, and information spread, particularly relevant for topics like vaccination, nutrition, or drug use.
   - Predicting the spread of health misinformation and devising strategies to counteract.
   - Sentiment analysis example.
   - Sentiment analysis review.
   - Monitoring and analysis of online data or mobile device data can help detect early signs of depression, anxiety, or other mental health issues. Example
7. Environmental Health:
   - Using sensors and other data sources, machine learning can help track pollution or other environmental hazards and correlate them with health outcomes.
   - Air pollution estimation

Our goals

- The overarching goal of this class is to introduce machine (or statistical) learning methods.
- Main topics:
  - Introduction to machine learning in public health
  - Data Preprocessing and Exploratory Data Analysis (EDA)
  - Supervised Learning for Disease Prediction and Classification
  - Unsupervised Learning for Public Health Data Analysis
- Potential goals of the analysis:
  - determine which predictors impact the outcome, or
  - predict the risk of disease (0/1) or value of a (continuous) biomarker.

Goals of an analysis (unsupervised)

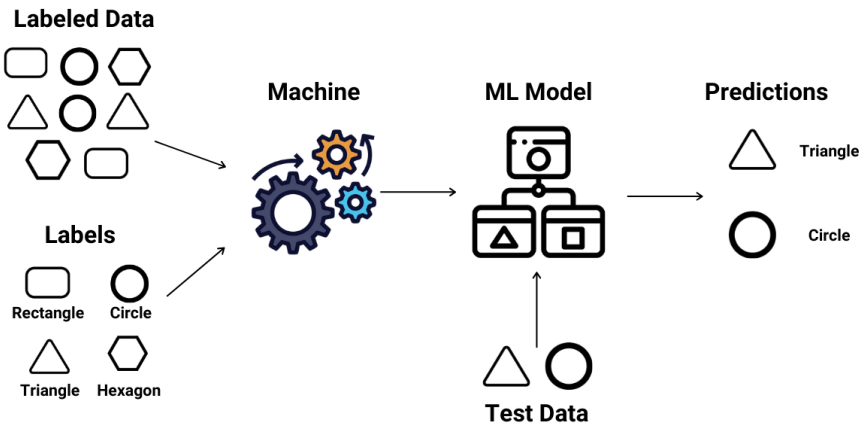
- **Descriptive**: search for trends, outliers, clusters, latent structures, multivariate distributions.
- Descriptive techniques lead to a field called **unsupervised learning**.


UNSUPERVISED LEARNING
Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.
Input Raw Data
DatabaseTown
Outputs
Interpretation
Processing
Algorithms
Unlabeled Data

Goals of an analysis (supervised)

- **Predictive**: Build models for classification, prediction, pattern recognition, and estimate predictive accuracy.
- Predictive models lead to **supervised learning** that we'll discuss throughout, and more so than unsupervised learning.

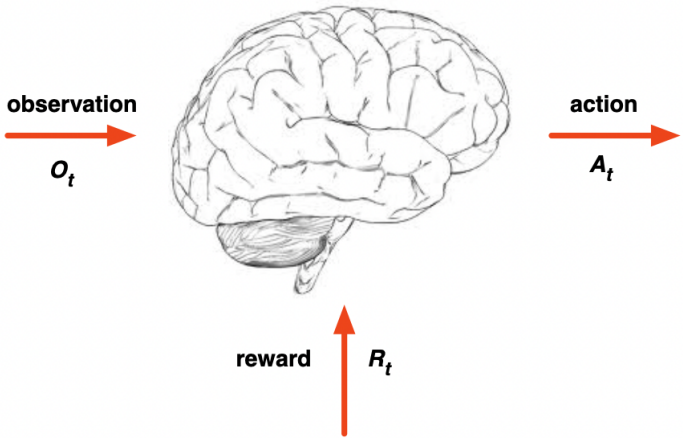
Supervised Learning



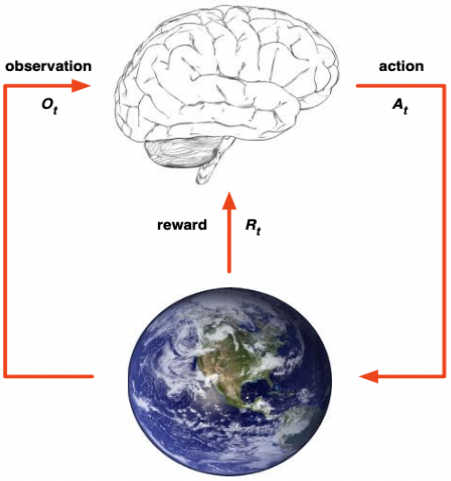
Reinforcement learning

- Reinforcement learning differs from unsupervised and supervised learning in that the model has to learn how to choose a **sequential series of decisions** to maximize a reward.
- The learner is not told which actions to take but instead must discover which actions yield the most reward by trying them.

Agent and Environment



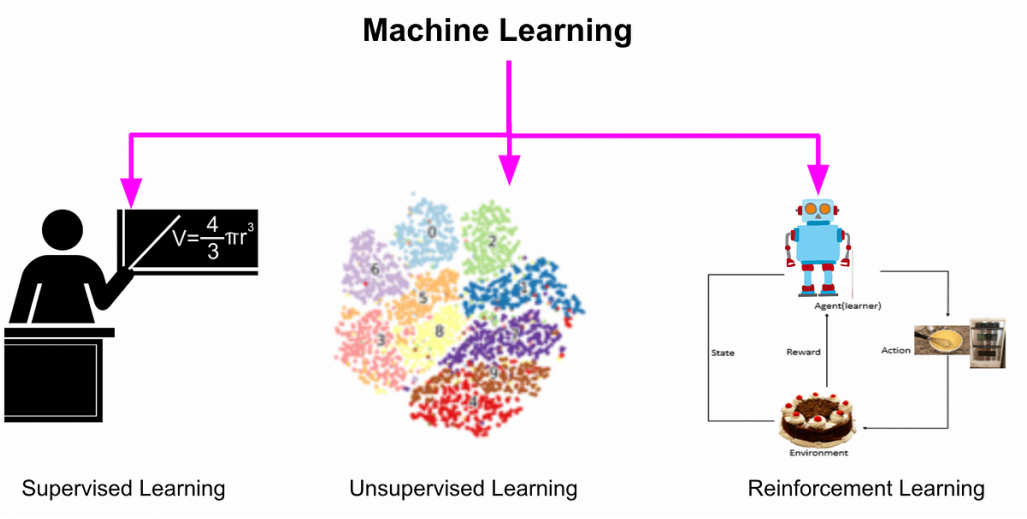
Agent and Environment



- At each step t the agent:
  - Executes action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at env. step

Reinforcement learning

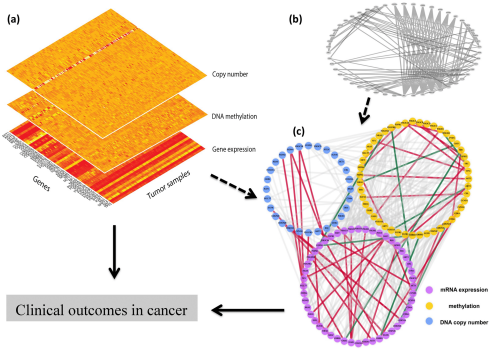
- In some cases, actions may affect not only the immediate reward. They affect the next situation and, through that, all subsequent rewards.
- **Trial-and-error search and delayed reward** are the two most important distinguishing features of reinforcement learning.

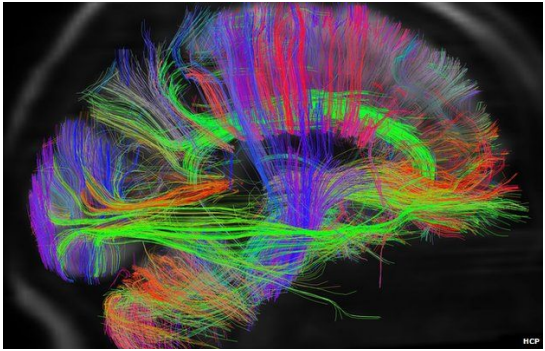

Machine Learning
$V=\frac{4}{3}\pi r^3$
Agent(learner)
State
Reward
Action
Environment
Supervised Learning
Unsupervised Learning
Reinforcement Learning

Data Types

- Data and computational power have become cheaper and more abundant than they were 10-20 years ago.
- Efficient analyses of complex data are more feasible.
- We will consistently discuss different data examples. In general, we'll put data into one or more of the following categories:
  - **Skinny data:** $p \leq 30$ as $n > 500$
  - **Wide data:** $p > 10,000$ with $n < 1,000$
  - **Large data:** $p = 100,000$ with $n = 10,000,000$
- These categories are useful for determining which is the best method to use; however, the most important category might be:
  - **Messy data:** multiple sources, coding systems, and data quality issues.

How large can p get?

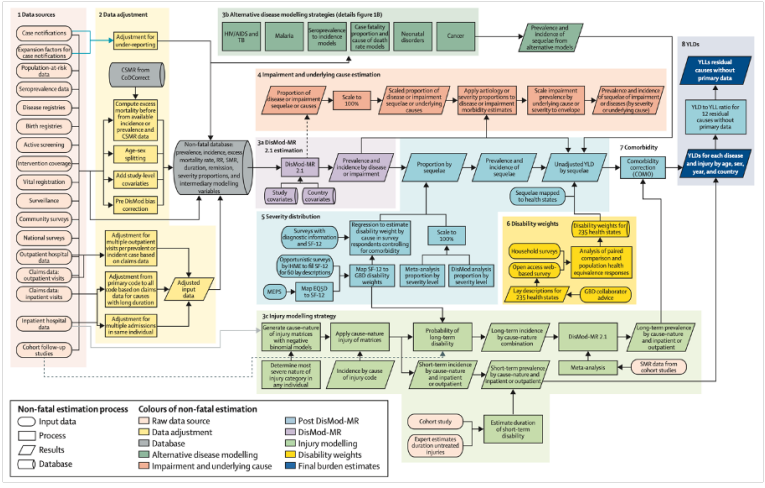




How large can p get?

- Genomic or neuroimaging data can have millions (10^6) of predictor variables
  - Not to mention genomic $\times$ neuroimaging interactions.
  - Electronic health records (EHR) data.
- There are some large datasets (UK Biobank $n \approx 5 \times 10^5$) but most are small (around 100–500).
- This is known as the $p \gg n$ setting.

Data Complexity/Messyness



Data Example

Cancer Cell Line Encyclopedia

- This database contains 136,488 datasets covering a total of 1,457 cancer cell lines.
- Pharmaceutical companies and academic researchers use the CCLE to screen potential anticancer drugs.
- For example, we have data measuring responses to 8 anticancer drugs. For each, we 18,988 human gene expression levels.
- The outcome Y measures the drug's effectiveness.
- The goal is to identify biomarkers for drug response and resistance, which can aid in developing targeted therapies.

Data Example

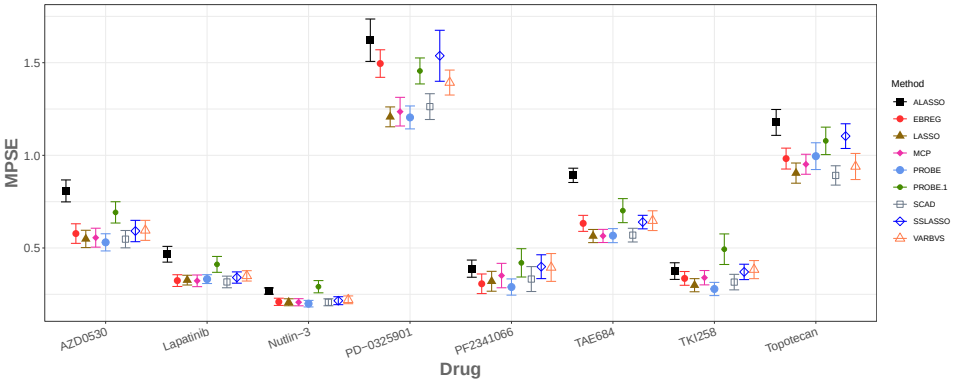


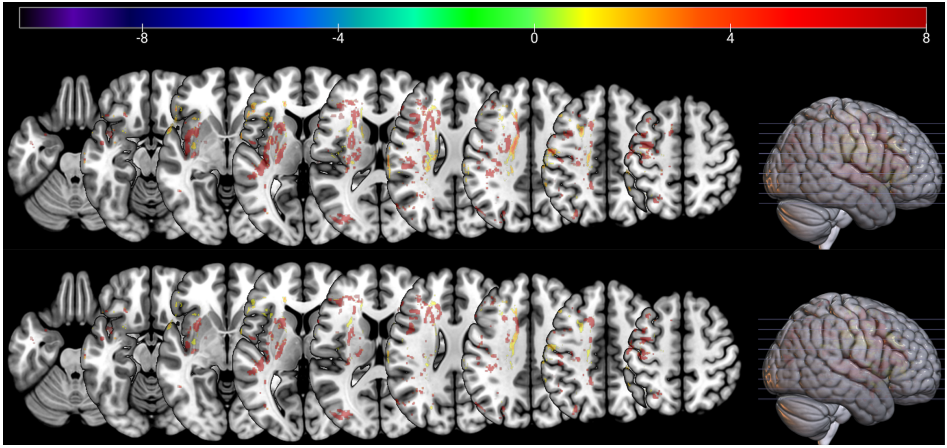
Figure: CV-MSPE: overall MSPE $\pm$ the estimated standard error (all based on 10-fold CV).

Data Example

POLAR (Predicting Outcomes of Language Rehabilitation in Aphasia) trial

- A total of 107 stroke patients with chronic aphasia (speech disorder) were randomized to one of two treatment arms.
- Neuroimaging data on where their stroke occurred is available on $\geq 5 \times 10^6$ voxels
- Main outcome is the Western Aphasia Battery (WAB).
- Goals of the study:
  1. see which treatment arm was the most effective,
  2. predict a person's WAB score based on their neuroimaging data,
  3. determine which areas of the brain with damage are related to WAB, and
  4. determine which areas of the brain with damage are related to treatment efficacy.

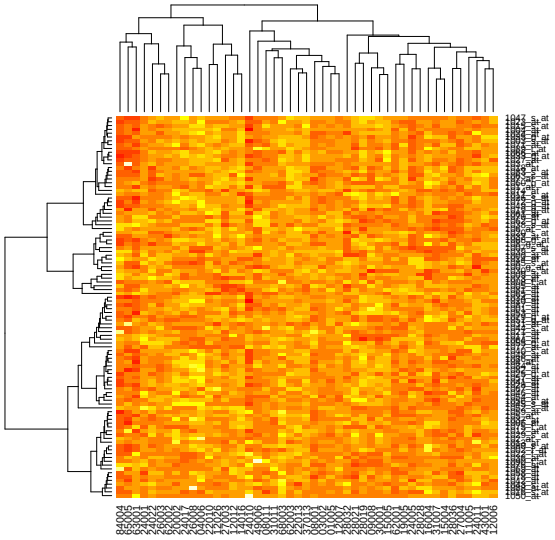
Areas of the brain related to WAB

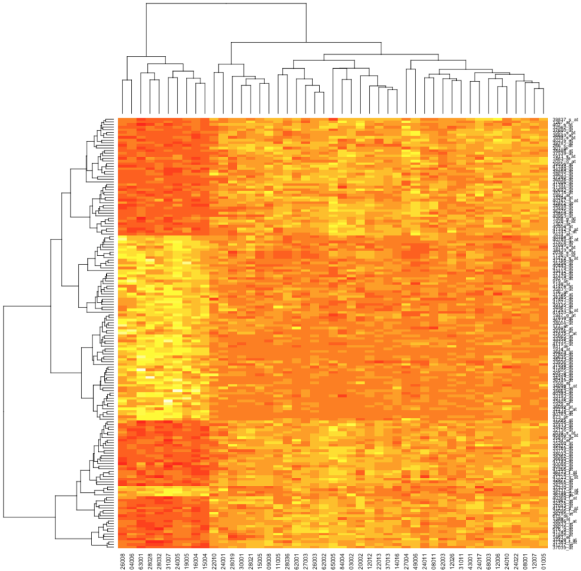


Data Example

Acute Lymphoblastic Leukemia Data from the Ritz Laboratory

- The data consist of $12,625$ microarrays from 128 different individuals with acute lymphoblastic leukemia (ALL)
  - This data is available in `R` via the `ALL` package
- Some additional covariates are available (sex, age, info about the diagnosis).
- ALL data enables the development of personalized treatment plans based on the genetic profile of an individual's leukemia cells, improving treatment efficacy and reducing side effects.
- An important step in this process is grouping ALL patients according to their genetic profile.

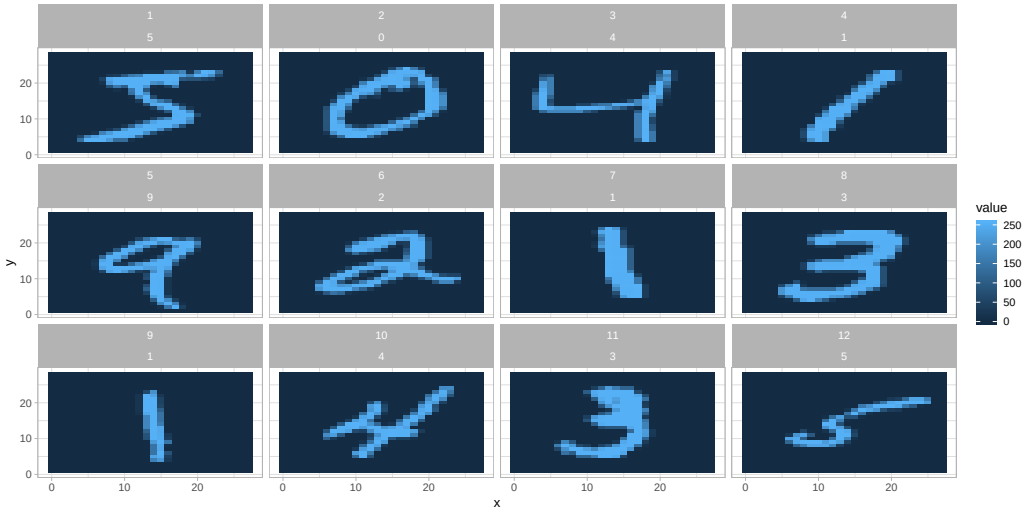




Data Example

Hand-written digits dataset

- This data consists of handwritten ZIP codes on envelopes from U.S. postal mail.
- The images are 16×16 eight-bit grayscale maps, with each pixel ranging in intensity from 0 to 255.
- The images have been normalized to have approximately the same size and orientation.
- The goal is to predict, from the 16×16 matrix of pixel intensities, the identity of each image $(0, 1, \ldots, 9)$ quickly and accurately.
- The dataset has more than 100,000 observations


1
5
2
0
3
4
4
1
5
9
6
2
7
1
8
3
9
1
10
4
11
3
12
5
y
x
0
10
20
value
250
200
150
100
50
0

Supervised learning introduction

- **Supervised learning** consists of trying to predict an outcome from several *features* or *independent variables*.
- The outcome could be continuous $\rightarrow$ regression.
- The outcome could be categorical $\rightarrow$ classification.

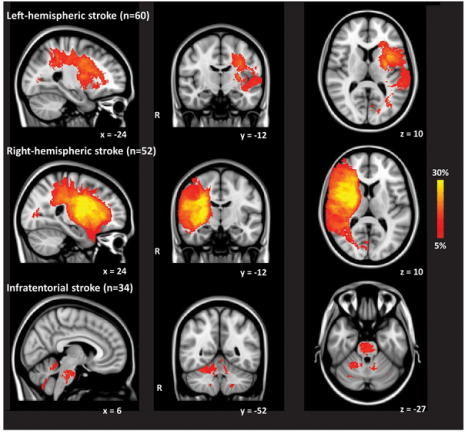
In general

- Y: outcome.
- $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$: independent or predictor variables.

Supervised learning example

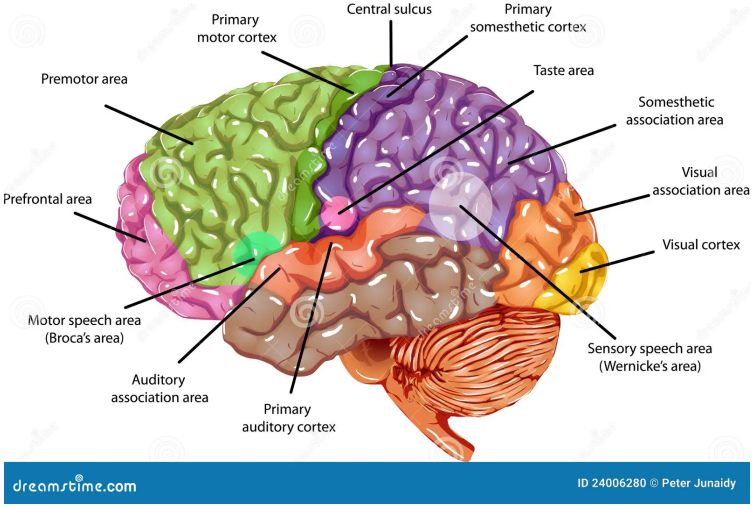
- Let's return to the POLAR study.
- All patients in the study have had a left-hemispheric stroke. We have brain images denoting where the stroke occurred.

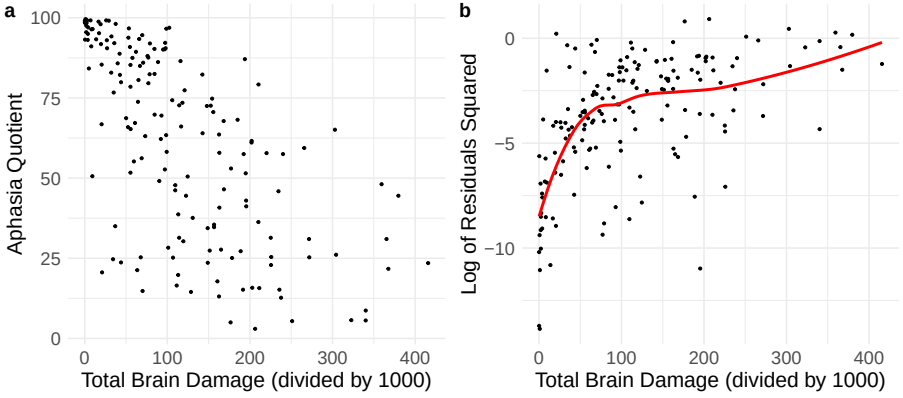
Schaapsmeerders, P., Tuladhar, ... & de Leeuw, F. E. (2015). *Lower ipsilateral hippocampal integrity after ischemic stroke in young adults: a long-term follow-up study.* PloS one, 10(10), e0139772.



Supervised learning example

- The outcome of interest is the subjects' Aphasia Quotient (AQ), a 0–100 score quantifying language impairment.
- This score is vital to understanding patients' treatment options.
- Collecting AQ is a cumbersome task, particularly for patients who have recently had a stroke.
- Consequently, it is of interest to develop models that can *predict* subjects' unknown AQ based on images of their brains.


Primary motor cortex
Central sulcus
Primary somesthetic cortex
Premotor area
Taste area
Somesthetic association area
Visual association area
Prefrontal area
Visual cortex
Motor speech area (Broca's area)
Sensory speech area (Wernicke's area)
Auditory association area
Primary auditory cortex
dreamstime.com
ID 24006280 © Peter Junaidy


a
Aphasia Quotient
100
75
50
25
0
0
100
200
300
400
Total Brain Damage (divided by 1000)
b
Log of Residuals Squared
0
−5
−10
0
100
200
300
400
Total Brain Damage (divided by 1000)

Multiple Linear Regression

- One dependent variable Y.
- p independent variables:
  - X_j = proportion of voxels damaged in Region Of Interest (ROI) j.
- The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- β_0: intercept;
- $\beta_j, j = 1, \cdots, p$: regression coefficients.
- $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$

Estimating $\boldsymbol{\beta}$

- $\boldsymbol{\beta}$ is estimated by minimizing

$$RSS(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{X}'\hat{\boldsymbol{\beta}})^2$$

- Denote the estimates of β_0 and β_j as $\hat{\beta}_0$ and $\hat{\beta}_j$.
- The fitted regression line (or the prediction/estimation equation) is then written as $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \beta_2 X_2 + \cdots + \hat{\beta}_p X_p$

What is Unsupervised Learning?

- **Goal:** Discover patterns or structure in data without using labeled outcomes.
- The model only sees input features $\mathbf{X}$, not a target variable Y.
- Useful for:
  - Grouping similar data (clustering)
  - Reducing data complexity (dimensionality reduction)
  - Feature learning and exploratory analysis

Clustering

- **Clustering** = grouping data points into clusters based on similarity.
- Common algorithm: **k-means clustering**
  - Assign each point to the nearest of k centroids
  - Update centroids by averaging points in each cluster
  - Repeat until assignments stop changing
- Other methods: hierarchical clustering, DBSCAN

Dimensionality Reduction

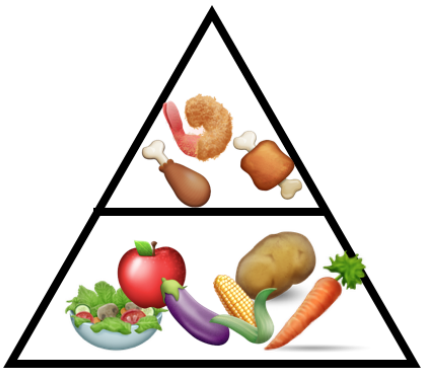
- **Goal:** Reduce the number of features while preserving important structure
- Example: **Principal Component Analysis (PCA)**
  - Find new axes (principal components) that capture the most variance
  - Project data onto the first few components to simplify
- Useful for visualization, noise reduction, and speeding up models

Clustering vs. Dimensionality Reduction

- **Clustering:** Group data (e.g., identify subgroups of patients)
- **Dimensionality Reduction:** Simplify data (e.g., reduce 100 features to 2)
- Can be used together:
  - Use PCA to reduce data to 2D, then apply k-means
- No right answer: outputs depend on the structure of the input data

Unsupervised Learning Example

- Consider data from the United States Department of Agriculture.
- This data contains the nutritional content of one serving of several food items (Parsley, Kale, Broccoli, Corn, Chicken, Beef, etc.).



	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

Figure: Factor loadings for the food example.

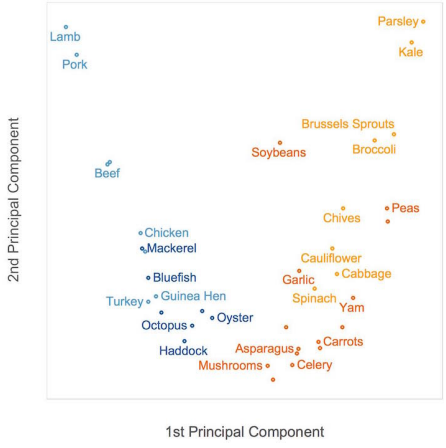


Figure: First and second PCs for the food example.

Case Studies

Disease Prediction

- AI-Enhanced Blood Test for Early Parkinson's Detection

Outbreak Detection

- Machine Learning-Based COVID-19 Outbreak Detection

Personalized Medicine

- MammaPrint: 70-Gene Signature for Breast Cancer Treatment Decisions

Why Linear Algebra for Machine Learning?

- Data tables ⇒ **matrices**; each row = observation, each column = feature.
- Core algorithms (e.g. linear regression, PCA, neural networks) = matrix operations.
- Understanding vectors/matrices helps diagnose issues like multicollinearity and rank deficiency.
- Dimensionality concepts (rank, eigenvalues) underpin dimension reduction.

Vectors: The Building Blocks

Definition

An n-dimensional **column vector** is

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

- Geometric view: a directed arrow in $\mathbb{R}^n$.
- Key operations

  $\boldsymbol{v} + \boldsymbol{w}, \quad c\,\boldsymbol{v}, \quad \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}^\top \boldsymbol{w}, \quad \|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^\top \boldsymbol{v}}.$
- In regression: the outcome y and each predictor column $\boldsymbol{x}_j$ are vectors.

Matrices: Organizing Data

Definition

An $n \times p$ **matrix** $\boldsymbol{X}$ stacks p column vectors:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

- Basic operations: matrix multiplication (AB), transpose ($\boldsymbol{X}^\top$), inverse ($\boldsymbol{X}^{-1}$, if square and full rank).
- **Rank**: number of independent columns; ties to multicollinearity.
- **Eigenvalues**/**Singular values**: measure variance explained; basis of PCA.

Connecting to Linear Regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

- $\boldsymbol{X}^\top \boldsymbol{X}$ is a $p \times p$ matrix capturing predictor relationships.
- Inversion requires $\boldsymbol{X}$ to have full column rank ($p \leq n$ and no perfect collinearity).
- High p relative to $n \Rightarrow$ need for penalisation (ridge, lasso) or dimension reduction (PCA).

Dimensionality: Why It Matters

- **Curse of Dimensionality**: distance metrics and volume behave counter-intuitively as p grows.
- Many ML algorithms scale as $\mathcal{O}(np^2)$ or $\mathcal{O}(p^3)$; large p is costly.
- Dimension reduction (e.g. PCA, t-SNE) projects data into a lower-rank subspace preserving key structure.
- In neuroimaging/genetics, thousands of predictors $\Rightarrow$ combine shrinkage + low-rank priors.

Key Takeaways

1. Vectors and matrices are the native language of modern ML algorithms.
2. Basic operations (dot product, matrix multiply) underpin model fitting.
3. Rank, eigenvalues, and singular values diagnose multicollinearity and guide dimension reduction.
4. A solid grasp of these ideas helps you choose, implement, and troubleshoot machine-learning methods in R.